Some Conceptual Issues in Observed-Score Equating

Wim J. van der Linden
CTB/McGraw-Hill

In spite of all of the technical progress in observed-score equating, several of the more conceptual aspects of the process still are not well understood. As a result, the equating literature struggles with rather complex criteria of equating, lack of a test-theoretic foundation, confusing terminology, and ad hoc analyses. A return to Lord’s foundational criterion of equity of equating, a derivation of the true equating transformation from it, and mainstream statistical treatment of the problem of estimating the transformation for various data-collection designs exist as a solution to the problem.

One of the oldest and most researched topics in educational measurement is observed-score equating. Issues associated with observed-score equating are not only theoretically challenging, but their solutions also are critical to the practical success of educational measurement. If we are unable to adjust the scores of examinees on one form of a test to make them indistinguishable from those on an alternate form, we should seriously wonder if they represent any measurement at all. Incomparable test scores are just numbers assigned to examinees—no quantitative inferences as to differences between their abilities or growth of knowledge can be made from them.

The key challenge in score equating has been clearly identified by Holland and Dorans (2006, p. 193). After they formulate the basic problem of comparing the scores of two different groups of examinees on two different test forms, they continue by stating:

In examining the distributions of the resulting scores, there are two, ever-present, factors that can influence the results, no matter how similar the score scales of the two tests appear. One is the relative difficulty of the two tests (which is what test scaling and equating is concerned about), and the other is the relative ability of the two groups of examinees on these tests (which is the confounding factor that should be eliminated in the linking process). In test linking, we are only interested in adjusting for differences in test characteristics and we wish to control for possible examinee differences in ability when making these adjustments.

I will adopt the same use of the difficulties of the different test forms as a pars pro toto for all the effects of items on the test scores. An equating thus is successful only to the extent that it neutralizes the effects of both the differences in difficulty between the test forms and the differences in ability between their examinees.

Early in the history of observed-score equating, both the literature and daily practice settled on the use of one general method of score equating: equipercentile equating. In essence, the method derives the score distributions of the two groups of examinees on the two forms for a common target population and then equates the scores in one distribution to those in the other. Suppose X is an existing test form and Y is
a new form developed as an alternative to it. We use $X$ and $Y$ to denote the scores on these two forms and $F(x)$ and $G(y)$ for the cumulative distribution functions (cdfs) of the scores for the target population, respectively. Following the tradition, we will refer to these two distributions as target distributions. Finally, let $x = \varphi(y)$ be the transformation that equates the scores on $Y$ back to those on $X$.

The equipercentile transformation is

$$x = \varphi(y) \quad (1)$$
$$= F^{-1}(G(y)).$$

The transformation can be derived using a common cumulative proportion $p$ of scores in the two target distributions as the criterion. Writing the inverses of the two cdfs as a function of $p$, we obtain

$$x = F^{-1}(p), \quad (2)$$

and

$$y = G^{-1}(p). \quad (3)$$

Elimination of the argument immediately gives us Equation 1.

All other transformations in the literature on observed-score equating are special versions of this transformation derived for specific equating designs or assumptions about the test scores. The three most common equating designs are based on test forms with (i) common items, (ii) common examinees, or (iii) randomly equivalent examinees. An example of a specific assumption is the one in which the target distributions for $X$ and $Y$ differ only in their means and standard deviations. The equipercentile transformation then takes a linear form, with different slope and intercept parameters for different additional assumptions. Also, modern applications of Equation 1 involve several technical steps to deal with issues such as irregularities in the shapes of the raw-score distributions, the general discreteness of test scores, and the derivation of the target distributions. In fact, solutions to most of these issues have dominated the recent literature, culminating in the sophisticated package of equating procedures known as kernel equating (von Davier, Holland, & Thayer, 2004). I deliberately omit all of these more technical issues from my discussion and focus on the conceptual aspects of the core idea of applying Equation 1 to target distributions for different test forms, which underlies all of current observed-score equating.

At first sight, the choice of Equation 1 seems convincing and it becomes even compelling if we notice its prominence in nearly every other area of statistics. Inverses of cumulative distribution functions (cdfs) are known as quantile functions, and because of Equations 2–3, the transformation in Equation 1 is more generally known as a quantile-quantile (Q-Q) transformation. One popular use of the transformation is in the form of a Q-Q plot used to compare the shape of an empirical distribution to its theoretical expectation (e.g., Wilk & Gnanadesikan, 1968). Every major statistical software package offers a variety of such plots. For instance, if we expect the shape of an empirical cdf, $G(\cdot)$, to be normal, the software chooses $F(\cdot)$ to be the normal cdf with the same mean and variance as $G(\cdot)$ and then plots Equation 1. The closer the curve to the identity line, the closer the shape of $G(\cdot)$ to normality.
Some Conceptual Issues in Observed-Score Equating

The principle at work here is that the curve that shows the differences between the cdfs of two distributions is the same as the one required to map one onto the other. Hence, the Q-Q transformation also can be used to transform a random variable with a given type of distribution into any other. For instance, users of the statistical language R who want to go from a chi-square variable with a given number of degrees of freedom to one with a gamma distribution for a given shape parameter \( \lambda \) can do so using the statement `qgamma(pchisq(x,df), \lambda)` which is just the representation of Equation 1 for these two distributions in the R language.

In fact, the same process explains why a random number generator only needs to sample the standard uniform distribution, \( U \sim [0, 1] \). This distribution has the identity function \( u = G(u) \) as its cdf. Substitution into Equation 1 gives us \( x = F^{-1}(u) \), which transforms a number \( u \) drawn from the standard uniform distribution to a random number \( x \) from any given distribution \( F(\cdot) \), a sampling technique known as inverse transform sampling.

It is interesting to note the implicit use of Equation 1 in the early days of psychological testing, when new test forms were not equated back to old forms but instead it sufficed to normalize their score distributions (van der Linden, 2000, p. 437). For example, if \( X \) and \( Y \) are two different forms of an intelligence test and \( \Phi(\cdot) \) denotes the cdf of the normal distribution with the mean of 100 and standard deviation of 15 typically adopted for them, the normalized versions of the scores \( X \) and \( Y \) are

\[
\varphi(x) = \Phi^{-1}(F(x)),
\]

and

\[
\varphi(y) = \Phi^{-1}(G(y)).
\]

As \( \varphi(X) \) and \( \varphi(Y) \) now are identically distributed, it holds that \( \Phi^{-1}(F(x)) = \Phi^{-1}(G(y)) \). Solving for \( x \) gives us the equation in Equation 1. Thus, for the same mean and standard deviation, normalizing the score distributions on two different forms of a test is the same thing as establishing an equipercentile relation directly between their original scores.

As the equipercentile transformation has such good ancestry and it keeps demonstrating its usefulness in numerous daily applications throughout the entire field of statistics, why doubt its use in observed-score equating? The examples in the following section show why.

A Few Embarrassing Examples

The first example is intended only to illustrate how the equipercentile transformation can be used to translate the scores on any variable to those on any other, even if the variables have no meaningful relationship at all. The first scores are the ages of the passengers on the 1912 maiden voyage of the Titanic recorded for 1,064 of them—that are in a data set freely downloadable from the Internet (`titanic3.csv`, January 22, 2013). The other scores are those of 60,729 Grade 8 students on a mathematics test from a recent large-scale assessment. The distributions of both scores are shown in Figure 1, where the mathematics test is labeled as Test Form X. I used Equation 1 to “equate” the Titanic scores to the scores on the mathematics test. The
Figure 1. Distributions of the age of the passengers on the Titanic and the observed scores of Grade 8 students on a mathematics test (Form X).

Figure 2. Equipercentile transformation from the age of the passengers on the Titanic to the observed scores of Grade 8 students on the mathematics test (Form X), along with the distribution of the “equated ages.”

The first panel in Figure 2 shows the transformation produced by the R package equate (Albano, 2012). (The same package was used in all other examples in this article.) The second panel shows the distribution of the “equated ages” of the Titanic passengers, which, except for discreteness, now has the same shape as the distribution of the mathematics scores of the Grade 8 students. Obviously, the two groups are entirely different, and the idea of a “target population” as a mixture of them would never make any sense. Besides, the empirical nature of both variables differs completely. In fact, age is a variable with an absolute zero and a physical unit that never requires any equating. But the example does demonstrate the power of the equipercentile transformation, even when it is meaningless to use it.
The second example makes me wonder if the equipercentile transformation does lead to any plausible score equating at all, even if the two forms share exactly the same variable and group of examinees. Test Form X in Figure 1 actually consisted of the odd items in a mathematics test that was twice as long. Its choice enables us to use the even items as a second form whose distribution is shown in the first panel of Figure 3 (Form Y). For both forms, the scores ranged from 0 to 28. An arbitrary split of a longer test into two half-length forms guarantees measurement of the same variable by both. Equally important, it creates an equivalent-group design that allows us to equate the scores on Y to X for the same group of 60,729 Grade 8 students that took the original test. Unlike the previous example, there was no need to be concerned about possible differences between the empirical nature of the two scores or the choice of a target population. The second panel of Figure 3 shows the equipercentile transformation that mapped the scores on Y back to those on X.

However, an equivalent-group design also gives us the bivariate distribution of the examinees’ observed scores on the two test forms, and therefore enables us to evaluate the equated scores on Y directly against the actual scores for the same examinees on X. Figure 4 shows three examples of the conditional distributions of the actual scores on X for the examinees with Y = 10, 15, and 20. It also shows the equated scores for the same three values of Y produced by the equating transformation in Figure 3: \( \varphi(10) = 12.3 \), \( \varphi(15) = 18.3 \), and \( \varphi(20) = 22.4 \). The obvious question is, what gives us the right to assume that these equated scores are correct? From the perspective of these three conditional distributions, they look pretty arbitrary. Actually, the use of the conditional means of X given Y = y seems much more defensible as a summary of the students’ actual scores on X than the equated scores. The latter were calculated from the marginal distributions on X and Y only; they completely ignored the information available in these conditional distributions. A plea for the use of these conditional means—that is, the regression function of X on y— as an equating method seems rather bold, though. The observed-score equating literature
Figure 4. Conditional distributions of the observed scores on Form X of the mathematics test given $Y = 10$, 15, and 20. The arrows indicate the equated scores $\varphi(10) = 12.3$, $\varphi(15) = 18.3$, and $\varphi(20) = 22.4$ at these values produced by the equipercentile transformation in Figure 3.

has a long tradition of ruling out regression as an equating method (e.g., Kolen & Brennan, 2004, Section 1.3.1).

As discussed later, the choice of a target population is one of the Achilles’ heels of observed-score equating. Intuitively, we do not want any equated score to depend on any population at all. A case in point is Figure 5, which in its first panel shows the distribution of $X$ to which we want to equate along with two alternative distributions of $Y$ from which we could equate. The same score of $X = 25$ leads to an equated score of $x_1 = \varphi(25) = 37.30$ when the distribution on the left is used, but $x_2 = \varphi(25) = 11.45$ for the distribution on the right. The difference between these two equated scores is huge. Admittedly, the three distributions are entirely fictitious.
Figure 5. Equating from two alternative populations for Form Y of a test to a single population for Form X. The left panel shows the three population distributions; the right panel the two alternative equipercentile transformations. An examinee with the score of $Y = 25$ receives an equated score equal to $x_1 = \phi(25) = 37.30$ or $x_2 = \phi(25) = 11.45$, depending on the population of Form Y (s)he is assigned. The arrows indicate the positions of the score $Y = 25$ (left panel) and the equated scores associated with it for the two alternative populations (right panel).

But the score of $X = 25$ is located right in the middle of the scale. And it is easy for a large-scale testing program to encounter groups of examinees on alternative administrations of the same test form with distributions differing as substantially as they do in Figure 5 (e.g., in school districts that differ in educational opportunity). My only intention is to show how dramatically the equated score for one examinee depends on the performances of all other examinees in the group to which (s)he happens to be assigned.

The next example illustrates the same problem of group dependency in more detail. All distributions in Figure 6 are for two test forms, X and Y, each consisting of 50 items randomly sampled from a previously operational pool that fitted the three-parameter logistic (3PL) response model. The two forms thus measure a common variable. The first two panels show the observed-score distributions on X and Y for examinees with ability level $\theta = -2$. (The observed-score distribution for a given $\theta$ under the 3PL model is known to be generalized or compound binomial and can be calculated directly from the item parameters.) As we have two regular observed-score distributions, the equipercentile transformation in Equation 1 can be used to map each score on Y onto the scale of X for any examinee with this ability level. The next two panels show the same two distributions for ability level $\theta = 2$. Again, we used the equipercentile transformation to map the scores on Y on the scale of X for the examinees at this higher level. Finally, we treated all examinees at these two ability levels as a single group. Assuming equal numbers of examinees at both levels—a choice only apparently innocent; see below—we obtained the two merged observed-score distributions in the last two panels of Figure 6 as well as another equipercentile
Figure 6. Observed-score distributions on Forms X and Y for examinees at $\theta = -2$ (row 1) and $\theta = 2$ (row 2), along with the observed-score distributions for a population consisting of equal numbers of examinees at these two ability levels (row 3).
transformation from the scores on Y to X. Figure 7 shows all three transformations. Each of them is different and returns a different equated score for each \( Y = y \). In statistical terms, the differences point at a problem of equating bias: the attempt to combine the observed-score distributions of examinees with two different levels of ability led to an equating transformation that necessarily compromised between the two underlying transformations. The closer it was to either, the farther away it was from the other. As a result, all equated scores were systematically off target and thus were biased.

I will now turn to an example of equating directly from the daily practice of educational testing. The equating was for two different forms of a biology test administered to high-school students as part of their exit exam. One form was used for the main administration, and the other was used as a retake examination. Both consisted of 28 multiple choice and 12 constructed-response items; 12 of these items were common. Both forms were constructed to be parallel in content and have minimal psychometric differences. The main form and the retake form were taken by 1,463 and 159 students, respectively. Their score distributions are shown in Figure 8. In order to guarantee fairness, the scores on the retake form were planned to be equated to those on the main form; hence the common items. As the number of retakers was especially low, linear equating was conducted—an option typically used for small samples because of the trade-off between an increase in bias against the typically
Figure 8. Observed-score distributions of the main administration and retake of a biology test in a high-school exit exam.

Figure 9. Linear equating transformation from the retake form back to the main form of the biology test in Figure 8.

greater gain in accuracy it offers (see the section on the trade-off below). The Tucker method (e.g., Kolen & Brennan, 2004, Section 4.1) yielded an equating line with a slope equal to 1.636 and an intercept equal to $-5.833$. As shown in Figure 9, a considerable portion of the equated scores for this transformation was completely out of range. The equated score of the student with the maximum score of $Y = 41$ among the retakers (see Figure 8) exceeded the maximum possible score of $X = 51$ on the main form by some 10 points. Obviously, reporting such equated scores would have created a revolt. Several of the retakers would have wondered how they could ever have failed the main test now that their current performance pointed at scores that are much higher than the maximum possible score. The reason for the anomaly,
Figure 10. Two different test forms with Guttman items. All items in Form X are located at \( \theta_1 \); all items in Form Y are at \( \theta_2 \).

of course, is the strong selection effect due to the pass-fail decision. Assuming that the two forms did not differ greatly, all retakers appeared to perform at the level of the lower part of the score distribution for the main test (Figure 8). Consequently, the equating method took the retake form to be much more difficult and compensated for it by producing higher equated scores. In a recent study, Puhan (2011) addressed the impact of the presence of repeaters on observed-score equating. His findings from an empirical study, which compared the results from the equating of different forms of a certification test with all repeaters included and removed, suggested negligible impact. The current example, however, involved the more serious case of a new form taken by repeaters only (except for one student). Its equating transformation was so unreasonable that it was never used.

The final example addresses a case in which equipercentile equating visibly leads to a dilemma. Suppose the two test forms consist of Guttman items. For ease of illustration, we assume that all \( n \) items in one of the two forms are located at \( \theta_1 \) and in the other at \( \theta_2 > \theta_1 \) (see Figure 10). Obviously, Y has an observed-score distribution with all scores at \( y = 0 \) or \( y = n \), and the same holds for the distribution on X. The dilemma exists because of the fact that the scores \( Y = 0 \) now have to be mapped onto two extremely different scores on X: (i) \( x = 0 \) for all examinees with \( \theta < \theta_1 \) and (ii) \( x = n \) for all examinees with \( \theta \geq \theta_1 \). No single equipercentile transformation from the scores on Y to X is able to create this mapping. Notice that the problem has nothing to do with population independence; we have not specified any population at all. Our first reaction might be to point to the unrealistic nature of the items: tests with pure Guttman items do not exist. But the dilemma actually is a specific version of the problem seen with the two test forms with the operational items in Figure 7, where for each possible score on Y we had to choose between as many different equated scores as ability levels in the group of examinees. The extreme shape of the Guttman response functions in the two forms only reduces the problem to the choice
between their two possible scores of 0 and \( n \). In fact, the example thus shows that the differences between the alternative transformations in Figure 7 can be expected to be much more extreme for test forms with more highly discriminating items and a larger difference in difficulty.

My conclusion from these examples is as follows:

**Conclusion 1:** Equipercentile transformations are Q-Q transformations. They can be used to equate the shape of any given score distribution to any other. They do not necessarily equate the observed scores on different test forms.

**EQUATING CRITERIA**

The literature offers several reviews of equating criteria (e.g., Harris & Crouse, 1993; Kolen & Brennan, 2004, Section 1.3; Morris, 1982; von Davier, Holland, & Thayer, 2004; Yen, 1983). One of the first was Lord’s (1980, Section 13.5), which consists of the following four criteria:

1. Unidimensional test forms measuring the same variable.
2. Equity, that is, for every \( \theta \) the equated scores, \( \varphi(Y) \) and the scores to which we equate, \( X \), must have the same conditional distribution.
3. Invariance across groups, which means that the transformation \( \varphi(y) \) must be the same regardless of the population used to determine it.
4. Symmetry, which is obtained only if the equating is independent of which test is labeled as X and which as Y.

Twenty years later, Dorans and Holland (2000) added another criterion, requiring both test forms to have equal reliability. At the same time, they changed the order in which the original criteria were presented: their new criterion of equal reliability has assumed the second position (directly after the one of measurement of the same variable), while the equity and group invariance criteria were moved to the bottom of their list. Also, a less statistical formulation of the equity criterion was adopted: “it ought to be a matter of indifference for an examinee to be tested by either one of two tests that have been equated” (p. 4). Finally, the notion of invariance across groups was replaced by population invariance. Since all changes were repeated in Holland and Dorans (2006, p. 194), I do not consider them as incidental and will return to their implications in my discussion below.

Generally, criteria are to help us distinguish between good and bad instances of something. The examples in the previous section were obvious cases of bad equating. If these criteria exclude them, we might still be in good shape. So let us analyze them a little further.

**MEASUREMENT OF THE SAME VARIABLE**

This criterion is self-evident; it has never been contested by anyone in the equating literature. Even though the Q-Q transformation can be used to give the score distributions of different groups of examinees on two different variables an identical shape, the result holds only temporarily. Any new examinees added to the study will
be ranked differently by the two forms, and the common shape of the distributions of the scores on X and the equated scores on Y immediately disappears (for more a fundamental discussion of the role of identical order in the definition of “same variable,” see van der Linden, 2000).

An important question therefore is how to establish measurement of the same variable in more subtle cases than our earlier “equating” of age to mathematical ability. Lord’s (1980) chapter on equating was written in the context of item response theory (IRT). This may explain why he referred to the criterion of the same variable rather casually in an introductory sentence and presented the other three as main criteria in a numbered list. In principle, establishing whether two test forms measure the same variable is not much of an issue in IRT—the only thing required is checking the joint fit of the response model to the items in both forms. The hypothesis of a common variable measured by the two forms is corroborated even more strongly if their cognitive structure can be modeled using the same Q-matrix (Xin & Zhang, in press).

Without IRT, identity of variables is more difficult to establish. The requirement that both test forms be constructed according to exactly the same set of specifications (Kolen & Brennan, 2004, Section 1.3.2) certainly is necessary, but it is not sufficient. Such forms still may yield scores that correlate less than satisfactorily, even after correction for attenuation; for instance, when some of the items in one of the forms appear to take more time than planned. When this happens, standardization of the time limit across the two forms becomes counterproductive. Rather than guaranteeing measurement of the same variable, it now creates different mixtures of ability and speed for both forms.

Equity

Lord’s definition of equity is statistical and unambiguous: If two unidimensional test forms X and Y measure the same variable θ, then for each possible value of θ, the distribution of the equated scores on one form must be identical to the distribution of the scores on the other form. Observe that the condition implies identical distributions of the observed scores on the two forms for each individual examinee. Lord’s definition thus gives us the exact statistical formulation of the requirement of indistinguishable scores referred to as the goal of any observed-score equating in the opening sentences of this article.

As will be argued later, the history of observed-score equating has been seriously hampered by the lack of an appropriate definition of what constitutes a true equating transformation—the transformation that delivers the perfectly equated scores that we try to estimate in our equating studies and against which we should evaluate our estimates. Lord’s equity criterion does allow us to derive this true transformation. From Conclusion 1, we know that the Q-Q transformation can be used to transform any given distribution into any other. So, let \( F_{X|\theta}(x) \) and \( G_{Y|\theta}(y) \) denote the cdfs of the distributions of the scores on X and Y for any given \( \theta \). As already demonstrated for the first two sets of plots in Figure 6, the Q-Q transformation can be used to map the latter onto the former:
Conclusion 2: For each given value of $\theta$ measured by $X$ and $Y$, the true equating transformation is

$$\varphi(y; \theta) = F_{X|\theta}^{-1}(G_{Y|\theta}(y)).$$

(6)

Morris (1982) adopted Lord’s definition of equity and provided a weaker version of it in the form of equality of the expected score on $X$ and the expected equated score on $Y$ for each $\theta$. Later, this version became known as “first-order equity,” and a criterion of second-order equity was introduced as the additional requirement of identical standard errors of measurement at each $\theta$ (e.g., Kolen & Brennan, 2004, Section 8.4.1). Both versions have been used to evaluate observed-score equating. For a recent study, see Tong and Kolen (2005). These studies thus used the means and standard deviations of the observed scores on $X$ and equated scores for $Y$ for each given $\theta$ only. The restriction would be inconsequential if their distributions were known to be from a location-scale family (e.g., normal distributions). But they are not; for instance, we know that observed-score distributions are much more skewed the closer they are to the ends of the scale. The transformation in Equation 6 can be used to evaluate the equating of observed scores with distributions of any form.

The alternative definition of equity in Dorans and Holland (2000), as a matter of indifference for an examinee to be tested by either one of two forms that are equated, looks like a serious step back from Lord’s. Not only does it miss its statistical precision, but, taken at face value, all it does seem to suggest is just asking examinees how indifferent they actually are. I also wonder why these authors demoted the criterion of equity to a position at the bottom of their list. I am aware of the pessimism created by Lord’s (1980) conclusion that, except for perfectly reliable or parallel test forms, exact equating was impossible. But his own reaction was much more pragmatic. In the same chapter, he ended his analysis by asking, “If you can’t... provide equity to everyone, what is the best next thing?” (p. 207). Setting a criterion aside because it cannot be perfectly met, rather than using it as a yardstick to find the next best thing, seems less appropriate. For a further discussion of this theme, I refer the interested reader to van der Linden (2011, Section 13.2).

Invariance Across Groups

I already have identified differences in ability between examinees in the two groups as one of the two main confounding factors in observed-score equating. The success of any equating depends critically on how effectively their effects are removed. Ever since Braun and Holland (1982), the standard solution is the introduction of two synthetic score distributions for $X$ and $Y$ that serve as the distributions for a target population for the equating. So far, we have simply used $F(x)$ and $G(y)$ to denote the cdfs of these two target distributions. In order to make the idea more explicit, let $P$ and $Q$ be the two groups of examinees that took $X$ and $Y$, respectively. We use $F_P(x)$ and $G_Q(y)$ for the cdfs of the empirical score distributions for these two groups on $X$ and $Y$, respectively. Likewise, $F_Q(x)$ and $G_P(y)$ are the cdfs of the score distributions that would have been obtained if, conversely, $Q$ had taken $X$ and $P$ had taken $Y$. 

262
The two target distributions on X and Y are defined by the following cdfs:

\[ F(x) = wF_P(x) + (1 - w)F_Q(x), \]

and

\[ G(y) = wG_P(x) + (1 - w)G_Q(x), \]

with \(0 \leq w \leq 1\). Of course, \(F_Q(x)\) and \(G_P(y)\) are unknown, and one of the achievements of kernel equating is its integrated approach to estimating \(F(x)\) and \(G(y)\) for all current data-collection designs.

A more fundamental problem than the estimation of these target distributions is their empirical status. Why should we ever be interested in these convex combinations of cdfs instead of those that were actually observed? And even if we accepted their necessity, what weights \(w\) should be chosen?

To begin with the latter, the recommendations for the choice of \(w\) in the literature include setting \(w = 0\) and \(1 - w = \) equal to the relative sizes of \(P\) and \(Q\), respectively, putting total weight on the group for which the scores are equated (here: setting \(w = 0\)), or giving equal weights to both groups (e.g., Kolen & Brennan, 2004, Section 4.1.5). I fail to see how the size of the group of examinees taking a test form could have anything to do with how its scores must be equated back to an earlier form. Also, putting total weight on one of the two groups of examinees seems ironic. It turns the whole idea of a unifying population that synthesizes the score distributions for two different groups into a word game that eventually takes us back to the situation we tried to fix.

The only possible choice consistent with the currently popular framework of observed-score equating seems \(w = 0.5\); that is, equal weights for \(P\) and \(Q\). This choice makes the equating independent of which test is labeled as X and Y and thus satisfies the criterion of symmetry below. As for any other choice of weights, to my knowledge, it has never been noticed that interchanging the labels of the two test forms requires a simultaneous relabeling of their two groups of examinees. Except for \(w = 0.5\), any interchange of \(P\) and \(Q\) in Equations 7 and 8 thus leads to two new target distributions, to another equating transformation than the inverse of the one from Y to X, and hence to violation of the criterion of symmetry.

But even for \(w = 0.5\), we fail to any attach any empirical meaning to the notion of a synthetic population. Score equating always takes place in a real-world context (e.g., the main and retake forms of the biology test in the example shown in Figures 8 and 9). How could the idea of a synthetic population ever solve this practical problem? The only admissible choice of \(w = 0.5\) would lead to an equating of the scores on Y for score distributions halfway between those for the students who passed and those who failed and had to retake the test. But the equating was meant to be for the latter. On the other hand, if we acknowledged this fact and put total weight on the distributions for the retakers, we still would fail to remove any of the effects of the (large) differences between the abilities of the passers and the retakers, the main reason why the equating was spoiled.

Actually, there is no need whatsoever to worry about group invariance. The true equating transformations in Equation 6 already have this feature! As they hold conditionally given each possible level of the common ability \(\theta\) measured by the two
forms, they are independent of the shape of any possible distribution over \( \theta \). An illustration of this independence already was given in the example in Figures 6 and 7. When we equated the observed scores at \( \theta = -2 \) and 2 separately, each of the two transformations was completely independent of the actual number of students at the \( \theta \) value. But when we merged the distributions for the two values, we had to account for the actual numbers of examinees at both of them! For the two plots in the last row of Figure 6, the arbitrary choice of equal numbers was made. If we had adopted different numbers, the plots would have been different and we would have obtained a different transformation.

From the same example, we already know that combining different ability levels leads to bias in the equated scores and therefore to loss of equity. The three criteria of equity, group independence, and unbiasedness of equated scores are thus equivalent—we always meet or fail all of them simultaneously (for a formal proof see van der Linden, 2011, Theorem 3). The common cause of their success or failure is one of the two main confounding factors in the quote in the introduction to this article: differences between the abilities of the examinees. But, unlike the current literature assumes, it is not just the difference between the ability distributions of the two different groups that take the two forms that counts but the differences between the abilities of the individual examinees in the entire equating study.

At first sight, this more serious type of confounding may seem daunting. How could we ever get rid of these individual differences? But the statistical technique of conditioning required for it is actually simpler and certainly less arbitrary than the construction of synthetic population distributions. As soon as we condition the observed-score distributions on \( X \) and \( Y \) on ability, this variable stops being a confounding factor. The only relevant questions thus are: (1) on what ability estimates should we condition? and (2) how much error will be incurred when using these estimates rather than the true ability levels? As we shall see later, the answers depend both on the data-collection design and the test-theory model adopted.

A popular technique of assessing the population invariance of an equating is to split the population into subpopulations of interest, perform the equating for each of them, and calculate the root mean square error of the differences between their equated scores as a function of the possible scores on \( Y \) (Dorans & Holland, 2000). In fact, this technique also involves conditioning but uses biographical variables as gender, language spoken at home, ethnicity, and so forth, as conditioning variables. In doing so, it leaves us with subpopulations of examinees still consisting of mixed levels of ability. As a result, it does not promote equity, leaves us with group dependency, and does not remove any of the bias in the equated scores.

I summarize our conclusion from these analyses as follows:

**Conclusion 3:** The true equating transformations in Equation 6 are already group-independent. The notion of a synthetic population is not necessary; actually, its use is one of the main sources of bias in observed-score equating.

**Symmetry**

This criterion has been motivated in two different ways. The more principled motivation points to the necessity of the use of the inverse transformation for the equating
Some Conceptual Issues in Observed-Score Equating

from X to Y. If this feature of symmetry did not hold, we would end up in an embarrassing situation: when equating back, we would be unable to reproduce the score on Y from which we originally equated.

The second reason is more pedagogic. A natural tendency of first-time students of equating is to confound the transformations used in linear equating with regression lines. As the two possible regression lines for the score on X and Y are not each other’s inverse, the criterion helps us to rule out regression analysis as a method of equating.

However, this second type of motivation is misleading in that it fails to make a distinction between the true equating transformation in Equation 6 and a possible estimate of it. The former is always symmetric: interchanging the labels of X and Y gives $\phi(x; \theta) = G_{Y|\theta}^{-1}(F_{X|\theta}(x))$, which is the inverse of Equation 6. At this level, the criterion is thus redundant. But at the level of a possible estimate, the criterion is too restrictive: In one of the next sections, I will argue that the problem of estimating an equating transformation is just an instance of the standard statistical problem of estimating an unknown quantity, complete with its evaluation criteria as (asymptotic) bias, accuracy, mean square error, and so forth. Imposing the feature of symmetry on its solution basically would restrict the class of possible estimators to plug-in estimators; that is, versions of Equation 6 with sample equivalents substituted for parameters. If the same restriction would have been adopted more generally in statistics, we never would have been allowed, for instance, to use $\hat{\sigma}^2 = \frac{1}{n-1} \sum (x - \bar{x})^2$ as an estimator of the population variance $\sigma^2$, even though it is known to have a smaller mean square error than the sample variance.

The distinction between true equating transformations and estimates sheds another light on the earlier reflection on the use of the regression function of Y given x as an equating function in the example in Figure 3. The only thing required to settle the issue is evaluating the regression function as an estimate of Equation 6 using standard statistical criteria. Actually, a first attempt already was made by Wiberg and van der Linden (2011), with mixed results for the lower end of the scale due to frequent guessing by the examinees. It remains to be seen if adjustments to deal with this problem of guessing might work.

**Conclusion 4:** True equating transformations are symmetric by definition. Requiring the same property to hold for estimated transformations is unnecessarily restrictive.

**Equal Reliability**

Frankly, I have been puzzled by the attention to this criterion. Dorans and Holland (2000), who introduced it, also appeared to be uncertain as to why it would be needed. Their main reason for its introduction sounded rather negative: “Without it, we could equate a reliable test to a single item, and this seems clearly mistaken” (p. 284). They also admitted that the criterion might be a “secondary consideration, not a fundamental requirement of test equating” (p. 284). Besides, they thought that “in addition to concerns about equal reliability of tests being equated, attention also should be focused on the amount of the reliability—more reliability is better for equating” (p. 285).
Both the idea of equal and more reliability might be remnants of Lord’s (1980, Section 13.2) conclusion that tests can be equated only if they are parallel or perfectly reliable. Equal reliability is a necessary condition for two tests to be parallel. And more reliability brings us closer to the ideal of perfect reliability.

But why introduce a criterion that is never used? Equipercentile equating ignores the reliabilities of the two forms entirely. The same holds for linear equating. Although the latter does involve assumptions about quantities that relate to the classical reliability coefficient, it only requires them to be equal for the same form in the two groups of examinees, not across the two forms. More importantly, the equating transformations in Equation 6 do not assume anything about the reliabilities of the two test forms. $F_{X|\theta}(x)$ and $G_{Y|\theta}(y)$ can represent any two score distributions; beyond being regular cdfs, nothing is assumed about their shapes. The former may be wide due to lack of reliability of the scores on X and the latter narrow, or just the other way around. Whatever the reliability of the two forms, the true transformations $\varphi(y; \theta) = F_{X|\theta}^{-1}(G_{Y|\theta}(y))$ automatically allow for any differences between them. Of course, these transformations have to be estimated and estimation error will propagate into the equated scores. But that is an entirely different issue than equal reliability of the two forms as a necessary requirement for their equating.

**Conclusion 5:** The criterion of equal reliability should be abandoned. The true equating transformation in Equation 6 automatically allows for any differences in reliability between the scores on X and Y.

*Any Examples Excluded?*

The set of equating criteria that have survived this analysis is even smaller than Lord’s. We definitely need measurement of a common variable and equity as criteria. These two criteria are *definitional*; together they define Equation 6 as the quantity of interest for observed-score equating. Two of the remaining criteria—group invariance and symmetry—are entirely redundant. They only describe features of equating that are necessary consequences of the first two. Finally, the criterion of reliability also is unnecessary. Of course, it is good for a test form to have high reliability. But we can always equate from any test form to any alternative form measuring the same variable.

So, does this smaller set of two criteria exclude any of the previous examples as invalid? It certainly does rule out the first example and is right in doing so; equating age to mathematical ability is just nonsense. But all other examples met the requirement of a common variable. They were only to demonstrate the consequences of violating the criterion of equity in the form of group-dependent, biased equated scores.

**Statistical Issues**

So far the discussion has focused mainly on the nature of the true transformations in Equation 6 and their relation to the equating criteria proposed in the literature. It now is time to discuss their estimation. In this section, I address only a few general statistical issues involved in this estimation. Details and examples follow once I have discussed the role of test theory in observed-score equating.
The problem of estimating a true equating transformation has the same elements as any other statistical estimation problem: (i) an unknown quantity of interest, (ii) data collected in a study designed to obtain information about it, and (iii) the need to translate the data into the best estimate of the quantity of interest. Therefore, there is no reason whatsoever to treat the current problem differently.

A useful analogy is the statistical approach to finding a good point estimator of an unknown parameter in a probability model. Its first step is to delineate the largest possible class of potential estimators. Once the class has been identified, the second step is to select a member that is best for the current data-collection design using standard statistical criteria. The common definition of the class of potential estimators is: “A point estimator is any function . . . of a sample; that is, any statistic is a point estimator” (Casella & Berger, 2002, Section 7.1). There are no initial restrictions on these functions or statistics. The only relevant issue is how well they do the job of estimating the parameter in terms of such criteria as unbiasedness, accuracy, etc.

We typically build test forms to have scores that increase in the common variable, so it seems safe to assume that they do not decrease in it. Consequently, equating transformations have to be monotonically nondecreasing. Obviously, the transformation should be from the range of all possible scores on X onto the range for Y. Finally, we follow the practical convention of continuizing test scores before equating them. These simple formal requirements lead to the following definition of an estimator of the equating transformation:

**Conclusion 6:** Any continuous, monotonically nondecreasing function \( x = \varphi(y) \) from the range of possible scores on Y onto the range for X is an estimator of the true transformation that equates the scores on the former to those on the latter.

It now is straightforward to define the equating error involved in the use of a given estimator \( \varphi(y) \) as its difference with the true transformation in Equation 6:

\[
e(y; \theta) = \varphi(y) - F_{X|\theta}^{-1}(G_{Y|\theta}(y)).
\] (9)

Observe that this definition leads to \( e(y; \theta) \) as a function of \( y \) for a given ability level, \( \theta \). The need to consider an entire error function makes sense; for each ability level, we meet examinees with different possible scores \( Y = y \), and for each \( y \) the use of \( \varphi(y) \) may imply a different error. [An alternative error definition is possible, based on how well, for a given \( \theta \), \( \varphi \) maps the distribution of the equated score, \( F_{\varphi(Y|\theta)}(x) \), onto the distribution of the observed score on X, \( F_{X|\theta}(x) \). Except for their metrics, the two definitions are entirely equivalent, though. The one in Equation 9 is easier to work with, but the alternative sometimes is more convenient (e.g., when we equate an adaptive test to a reference test); for details, see van der Linden (2006a, p. 360).]

The equating bias involved in the use of \( \varphi(y) \) as an estimator of Equation 6 is defined as

\[
\text{Bias}(y; \theta) = E(\varphi(y)) - F_{X|\theta}^{-1}(G_{Y|\theta}(y)).
\] (10)

while the standard error of equating is

\[
\text{SEE}(y; \theta) = E[\varphi(y) - E(\varphi(y))]^{1/2}.
\] (11)
If it is desirable to allow for the trade-off between these two measures in observed-score equating, the mean square error

\[
\text{MSE}(y; \theta) = \mathcal{E}\left[\varphi(y) - F^{-1}_{X|\theta}(G_{Y|\theta}(y))\right]^2 \tag{12}
\]

or its square root are helpful. Again, notice that each of these measures is a function of \(y\) for a given ability level, \(\theta\).

I have provided general definitions in Equations 10–12 that do involve taking expectations. This makes them appropriate for cases where the definition of \(\varphi(y)\) involves randomness. In the current context, an obvious case is estimation of \(\theta\) from the random response vectors of the examinees, where the necessity to take an expectation follows from the fact that the same \(Y = y\) can be the result of different vectors. Another example is equating from an adaptive test (van der Linden, 2006b).

But when \(\varphi(y)\) is fixed, the expectation operator must be dropped. The bias in Equation 10 then boils down to the difference between two different mathematical functions on the same range of possible scores on \(Y\). For each \(\theta\), the true equating transformation always assigns each examinee with the same \(Y = y\) to the same equated score, and \(\varphi(y)\) now does the same (but possibly to a different equated score, of course). Also, the SEE function becomes equal to zero for all \(y\). If there is no randomness in \(\varphi(y)\), all remaining error is systematic. For the same reason, the MSE function reduces to the square of the bias function. Each of these consequences of the absence of randomness makes perfect sense.

**Alternative Definitions of True Equating**

The equating literature shows two alternative definitions of true equating and, therefore, equating error. The first by Kolen and Brennan (2004, Section 7.1) is more explicit but less widely used. Their definition is in the same vein as the classical test-theory definition of a true test score. It assumes replications of the equating procedure across random samples from a population of examinees. For each sample, the transformation actually used is an estimate, \(\hat{\varphi}(y)\). These authors then define the true equating transformation as the expectation of \(\hat{\varphi}(y)\) across replications; that is, \(\mathcal{E}(\hat{\varphi}(y))\). Subsequently, equating error is defined as

\[
\hat{\varphi}(y) - \mathcal{E}(\hat{\varphi}(y)), \tag{13}
\]

while the expected value of its square is used to define the standard error of equating.

This definition of true equating seems inappropriate. Rather than defining a true transformation as the one that produces equated scores completely indistinguishable from the ones to which we equate, which we try to estimate, and against which we should evaluate our estimates, it accepts the expectation of any transformation that happens to be used as “true.” If we would follow the interest of current observed-score equating, which is mainly in the standard error of equating and tends to ignore the presence of equating bias, the goal of finding the best members of the class of all possible functions \(\varphi(y)\) as an estimate of \(\mathcal{E}(\hat{\varphi}(y))\) is bound to lead to degenerate cases of score equating. One example is always giving each examinee on \(Y\) its own
score as equated score; that is, using
\[ \varphi(y) = y. \]  
(14)

Another is to always give all of them the same arbitrary number \( c \) as equated score:
\[ \varphi(y) = c. \]  
(15)

In both cases, there is never any equating error according to Equation 13, and the standard error of equating thus is always equal to zero.

The second definition of the true equating transformation is the equipercentile transformation for a population of examinees adopted in the equating study. It seems more much accepted because it shares the idea of the assumption of a population with almost every other notion in the observed-score equating literature. But we hardly ever randomly sample any examinees from a population in educational measurement. The only exception that comes to mind is survey-style assessments, such as those in NAEP and PISA, with their primary interest in estimating (inter)national distributions of educational achievements.

It seems even harder to reconcile the notion of a synthetic population with any actual random sampling of examinees. Its definition in Equations 7 and 8 holds only for a stratified random sampling plan, with \( P \) and \( Q \) as strata from which each examinee is drawn with probabilities \( w \) and \( 1 - w \), respectively. I have never seen this type of sampling plan used in any equating study. Its examinees typically are assigned to test forms by school administrators on the basis of some objective criterion, they sign up for a form (for instance, as in admission testing), or they have to take a new form as the result of a poor performance on a previous form. Each of these options typically implies a probability of getting the other form rather than the one actually taken equal to zero, instead of either the same \( w \) or \( 1 - w \), \( 0 < w < 1 \), for all examinees. For instance, how could the two groups of students in the example in Figures 8 and 9—those that passed and failed the main test—ever have been produced by stratified random sampling as in Equations 7 and 8?

The usual escape from this type of criticism is to claim that the population we should be interested in does not need to be real as long as it is “intended.” Hence, it would be enough to establish a conceptual relationship between the actual examinees and this intended population that somehow involved the idea of random sampling from the latter. In a recent study, Kim, von Davier, and Haberman (2011) formulated this point as follows: “Equating quality depends on the relations between available equating samples and the intended population of examinees” (p. 112). But what kind of relations? And why would a Platonic world created by score equators be more important than the real world in which the examinees live?

The definition of the true equating transformation as an equipercentile transformation for an intended population sometimes is presented in more sophisticated language. Lord (1982), in the derivation of his standard error of observed-score equating, departs from the observation that the equipercentile transformation in observed-score equating follows from the requirement that the equated scores \( \varphi(Y) \) and \( X \) be identically distributed “for some specified population.” The standard error is necessary only to evaluate the effects of sampling fluctuations. He then reintroduces this idea of a true equating in statistical language by asserting that for
the equipercentile transformation, \( \varphi(Y) \) and \( X \) are always “asymptotically equated” (p. 165). But if more examinees automatically get us closer to our ideal, it no longer matters whether or not the population is operationally defined—the asymptotic argument discharges us of the necessity to become more specific.

The same asymptotic argument entails two more statistical consequences: First, neither does it suggest any need to be concerned about possible bias. Any existing bias automatically disappears. In fact, the same article more explicitly claims the only possible bias to be due to one missing observation in the calculation of the equipercentile transformation (Lord, 1982, p. 165), an entirely negligible effect. Second, the standard error of equating also decreases with the numbers of examinees. Indeed, Lord’s expression for this standard error of equating critically depends on the numbers of examinees that take \( X \) and \( Y \). Both figure in its denominator; if they increase, the standard error vanishes (Lord, 1982, Equation 8).

But this asymptotic argument in the numbers of examinees does not hold generally, and neither does any of its consequences. In fact, Lord’s (1980, Chapter 13) earlier analysis of observed-score equating begins with a discussion of a case that immediately shows its gratuitousness: the equating of infallible measures. The analysis rests on the observation that for two perfectly reliable test forms measuring the same variable, any group of examinees is ranked identically by their true scores on these forms. However, if this holds for any group, it holds for any two groups. But then, except for discreteness, any group of examinees always produces exactly the same equipercentile transformation between the two forms as any population from which it is supposed to be sampled. In addition, we already know that with increasing test length, test reliability increases and the observed scores converge to their true scores. It follows that for any equivalent-group design, the correct asymptotic argument must be based on the number of items in the test and not the number of examinees that take it.

The examples presented later in this article reveal the same behavior of equating error as a function of test length or reliability. Hence, the conclusion:

**Conclusion 7:** Equating error typically vanishes if the length of the test forms or their reliability increases, not necessarily with an increase in the numbers of examinees.

For the same reason, the use of resampling methods to estimate standard errors of equating, such as the currently popular bootstrap, is potentially misleading. Having a computer resample, a group of examinees easily reinforces the impression that we are dealing with sampling error. But it does not mean anything beyond the current groups of examinees unless they actually were sampled from an operationally defined population.

**Bias–Accuracy Trade-Off**

Both alternative definitions of a true equating transformation suggest that we only need to be concerned with the accuracy of observed-score equating. As for the possibility of bias in equated scores, Equation 13 declares them free of it by definition, while it is entirely negligible according to the population definition of true equating.
However, all of the earlier examples in Figures 3–10 did point at serious equating bias.

Generally, minimization of a standard error of estimation is not difficult, as long as we ignore the price to be paid in the form of increased bias. The two estimators in Equations 14 and 15 are clear examples of this trade-off; both of them deliver perfectly accurate estimates of equated scores but their bias is totally unacceptable. The same bias–accuracy trade-off shows up in any estimation problem; it therefore is one of the central themes of statistics. But in spite of its universal existence, the trade-off has frequently been ignored in the observed-score equating literature.

An exception is the recent study by Kim, von Davier, and Haberman (2008, 2011). These authors evaluated a new equating transformation in the form of a weighted combination of the identity transformation in Equation 14 and the equipercentile transformation in Equation 1 against use of the latter only. Obviously, more weight on the former increases the accuracy of the new estimator but also its bias, whereas the opposite occurs for more weight on the latter. The authors’ goal was to find the optimal compromise between bias and accuracy in terms of the mean square error of their equating. The study is refreshing in that, in the spirit of Conclusion 6, it dared to challenge the traditional plug-in estimator of the equipercentile estimator by evaluating an alternative for it. It also used standard statistical criteria for its evaluation (although still defined for a population definition of true equating).

A rather compelling philosophical argument makes us believe that the evaluation of bias in observed-score equating actually should be front and center. The only reason for equating the scores on a new form of a test to an old form is to avoid the bias incurred if we would ignore the differences between the forms and treat their scores as equivalent. As a matter of fact, equating transformations are measures precisely of this potential bias! It would be embarrassing if our estimates of them did not do the job they were supposed to do.

**Conclusion 8:** Equating is to remove bias from test scores. We should always check to see if the equating has left any of the bias or accidentally added more to it.

**Test-Theory Issues**

Frequent use of terms like “test,” “item,” and “scores” suggests an equating literature well founded in test theory. But actually, it has little to do with theory. We already noticed that in spite of the adoption of the criterion of equal reliability, none of the current equating procedures depends on it. The same holds for the notion of measurement error. The typical equating report ignores its existence completely; its standard error is only to evaluate the impact of “sampling fluctuations.” Rather than an application of a theory that builds on a psychometric model, equipercentile equating appears to be just an algorithm that applies a Q-Q transformation to numbers produced by examinees assumed to be obtained by stratified random sampling. A similar observation holds for the Tucker method of linear equating for an anchor-test design. Its only assumptions are about the regression lines for the two test forms on the anchor test; nothing is assumed about the nature of their scores. Conversely, and equally significant, Lord and Novick’s (1968) standard treatment of classical test
van der Linden

theory does not contain any extension that deals with the problem of observed-score equating; in fact, it never even refers to it.

The only classical test-theory exception we are aware of is Levine’s linear method for an anchor-test design, which adopts its model of an observed score as the sum of a true score and error for the two forms and the anchor test in the equating, and combines the assumption with Tucker’s to derive equations for the slope and intercept of the equating line. IRT observed-score and true-score equating were proposed by Lord (e.g., 1980), not surprisingly as attempts to get closer to the ideal of equitable equating. I will discuss the relationship between IRT and score equating later.

Kernel equating, as summarized in von Davier, Holland, and Thayer (2004), takes observed-score equating even further away from test theory. Although it deserves much credit for integrating most of current observed-score equating in a rather sophisticated framework, it reduces linear equating to the choice of two bandwidth parameters in an observed-score smoothing procedure. In doing so, it severs the only tie equating had with classical test theory through Levine’s linear equating. Neither does it include any IRT-based equating. Inspection of the subject index in von Davier, Holland, and Thayer (2004) is revealing: the term “reliability” is used three times in the entire book, “measurement error” never.

However, we need test theory to explain the nature of test scores and validate our assumptions about them. Without it, we just equate blind numbers and are prone to making errors. Several such errors already have been met, for example, in the form of the adoption of unnecessary equating criteria, an incorrect choice of the distributions required to equate scores, and the idea of examinee sampling as the only source of equating error. A more complete list is offered by the numbered conclusions in this article.

A full discussion of the potential contributions of the various test-theory models to observed-score equating is beyond the current focus on its more conceptual aspects. But one assumption underlying every model is of critical importance—the random nature of the responses of examinees to test items. The assumption reflects our common experience that performances on replicated tasks tend to vary. For some types of tasks, for example, those involving motor, perceptual, or decision skills, exact replication is possible. The literature on experimental psychology offers numerous examples of the variation in the performance associated with them (for instance, in the experiments conducted to demonstrate the existence of a speed–accuracy trade-off; e.g., Luce, 1986). For test items with more cognitive tasks, the effects of retention and learning typically prevent exact replication. But the use of items that are close to each other suggests the same random behavior. And the idea of randomness also has permeated the language in which examinees talk about their testing experiences (for instance, when they refer to a failure as “bad luck,” or a retake as a “new chance.”)

Whether the randomness is modeled more specifically (as in IRT) or just assumed (as in the classical test model), it forms the foundation of all of test theory. Educational measurement texts inspired by classical test theory generally refer to this randomness as “measurement error.” But in its basic model, the assumption of random test scores comes first, whereupon measurement error is defined as the difference with their expected values (Novick, 1966).
Some Conceptual Issues in Observed-Score Equating

Statistically, test scores therefore have to be represented by random variables; that is, variables with a probability distribution over their range of possible values. As the assumption of randomness is for each individual examinee, we have as many random variables for a given test form as examinees, each with a possibly different distribution. During an administration of the form, the only things actually observed are realizations of these variables in the form of one of their possible values. In this article, I have kept the status of the observed scores on the two forms explicit by treating them as distinct random variables $X$ and $Y$ for each examinee, while their observed values were denoted as realizations $X = x$ and $Y = y$.

It is exactly at this point that the literature on observed-score equating appears to go wrong. Although it does pay lip service to such notions as measurement error, unreliability, true scores, and so forth, it subsequently treats observed scores just as fixed constants (that is, from the perspective of test theory; we will become more specific below). We have no idea of how this tradition began, but it is might very well have been reinforced simply by an unfortunate choice of terminology in educational testing: The term “observed score” suggests something that already has been observed for an examinee rather than the entire set of possible values the score can take along with their probabilities of occurrence. The idea that an “observed score” itself is never observed just sounds awkward. Likewise, at the level of a group of examinees, the terminology becomes even more critical. We now have to distinguish between an observed-score distribution (probability distribution of the possible scores for an individual examinee), a distribution of observed scores for a group of examinees (mixture of the observed-score distributions of all of its members), and a distribution of realized observed scores for a group of examinees.

A test theorist is always aware of the background of the entire distribution of an observed score against which its realization must be evaluated. Notions such as test reliability, standard error of measurement, and correction for attenuation are meaningless without it.

It is against the same background that we have to equate observed scores from a new form $Y$ to an old form $X$. As we now have two different random variables $X$ and $Y$ for the observed scores on these forms for every examinee, we need a different transformation to map the distribution of $Y$ onto that of $X$ for each of them. As always, a Q-Q transformation does this job. For a given examinee, the actual equating procedure then exists of the use of the appropriate transformation to assign his/her realization $y$ of $Y$ (“observed value of the observed score on $Y$”) to the equivalent realization $x$ of $X$ (“equated score”). In practice, however, we need not be quite so specific. All test-theory models treat examinees with the same level of ability on a test measuring a unidimensional ability as exchangeable. Consequently, the same transformation can be used for each of them—a fact on which our definition of the true equating transformation in Equation 6 already capitalized, and which, as we will see below, enables us to pool their data to estimate the transformation for some of the equating designs. But, of course, examinees with the same level of ability do not necessarily produce the same realizations for their observed scores; they therefore may get different equated scores.

As already alluded to, the issue of the random nature of observed scores in traditional score equating actually is somewhat more complicated. Although from the
perspective of test theory, traditional equating treats observed scores as fixed constants, at the same time, it treats examinees as if they were randomly sampled from a population. The assumption of random sampling actually does involve the adoption of a distinct random variable for the observed score of each examinee on a given test form. But—and this is essential—the statistical model associated with it is that of independent, identically distributed (iid) random variables for all examinees. For an equating problem with two forms X and Y, the assumption amounts to iid variables with the distributions in Equations 7 and 8, respectively. Models with iid random variables are useful when educational tests are used to survey a population of students—not if the interest is in their individual scores.

**Conclusion 9:** Traditional observed-score equating operates under the statistical assumption of iid random variables for each test form, typical of survey research. It should operate under the test-theoretic model of ability-dependent random variables.

**Linguistic Issues**

We use words to communicate meaning, but what we actually intend to communicate may get distorted by our very choice of words, especially if they are used in other contexts too. As we just saw, the use of the term “observed score” in educational testing is a case in point. The same holds for “reliability.” A person who steals a wallet every time (s)he sees one is perfectly reliable according to the definition of the term in test theory—a qualification for which we should not count on too much support in our daily conversations, though.

Along the same lines, the adoption of “equipercentile equating” instead of “quantile equating” has been a missed opportunity. Not only would it have avoided the awkward impression that in observed-score equating distributions are always divided into a hundred equal portions before we equate them, it also would have enabled the field to profit from other areas of statistics where quantile transformations have been in use for a long time. A comparable comment holds for the replacement of Lord’s (1980) “group invariance” with “population invariance.”

A striking aspect of the equating literature is its rich vocabulary of terms used to describe attempts to link scores on different test forms to each other. Among the first to organize the vocabulary were Linn (1993) and Mislevy (1992), whose classifications consisted of linking procedures described as “equating,” “calibration,” “prediction,” and “moderation.” A more recent review of the equating literature added “scale alignment,” “battery scaling,” “anchor scaling,” and the use of “concordances” (Holland & Dorans, 2006). A similarly rich vocabulary has been developed to describe the results of all these different linking procedures. Examples are descriptions of linked scores as “equivalent,” “equitable,” “indistinguishable,” “interchangeable,” and “comparable” with the scores to which they are linked.

Each research tradition goes through an initial stage of conceptual development in which new ideas flourish and different terminologies emerge. But then Occam’s razor takes over, and the growth of knowledge essentially exists in the capability to explain more with fewer concepts. The developments in observed-score equating appear to follow an entirely different pattern, though. It is this author’s impression that the difference might be due to the pessimism created by Lord’s (1980) finding
that equity requires either perfectly reliable or parallel test forms. As already noted, Lord’s own reaction was more pragmatic; he did not want the perfect to become his enemy, and he moved on to find what would be next best. But somehow the idea that equating is either impossible (perfectly reliable forms) or unnecessary (parallel forms) got deeply ingrained in our collective memory.

My explanation has been motivated by the fact that, in spite of all terminological differences, each of these different types of linking uses nearly exactly the same equating designs and method as the equipercentile equating discussed in this article. The entire vocabulary therefore looks more like a linguistic solution designed to salvage the practice of observed-score equating that existed before Lord’s monograph rather than a classification of truly different methods. It seems as if the collective reaction to Lord’s analysis has been: Okay, score equating might be impossible or unnecessary, but if we call it “scale alignment” or a “concordance” and no longer refer to the results as “equated scores,” we are still able to move on. The following quote, which addresses the impact of a target population $T$ on equating methods with exactly the same calculations, perfectly illustrates the current use of different terms for the same method and results: “The equipercentile function is used for equating, for forming concordances, and for vertical scaling, battery scaling, and calibration. For equating, we expect the influence of $T$ to be small or negligible, in that case, we will call the scores equivalent or interchangeable. In other cases, $T$ can have a substantial effect in which case we will call the scores comparable in $T$” (Holland & Dorans, 2006, p. 202).

This use of unnecessary terms is confusing and impossible to explain to outsiders to the field of educational measurement. Our discussion of equating criteria has left only two serious criteria—measurement of a common variable and equity—which suggests the necessity of only two different terms to describe any attempt to link the scores on different test forms: prediction and equating. If two forms measure a different variable, we can still predict the scores on either from those on the other. This does not mean that our predictions are perfect. But this is exactly why we ought to report appropriate measures of prediction bias and accuracy along with them. If both forms measure the same variable, we can equate their scores. Again, this does not mean that the equated scores will be perfect, but the measures of equating bias and accuracy that should be reported along with them are precisely to convey this message.

Actually, true scores and ability parameters are “impossible” as well, for exactly the same reason as in Lord’s (1980) analysis of score equating: their observation requires perfect reliability. But this has never been a reason to give these notions up or talk about estimates of them in multiple languages. We are only required to estimate them in the best possible way and report our estimates along with measures of bias and accuracy to show how close we were able to get.

Conclusion 10: The terminology currently used in observed-score equating is confusing. We need only to make a distinction between prediction (measurement of different variables) and equating (measurement of a common variable) and use standard statistical measures to report the quality of our predictions and equatings.
Fundamental Problem

All previous issues go back to the same fundamental problem: Any equating problem always has two confounding factors—different difficulties of the two test forms and different abilities of the two groups of examinees that take them. It is impossible to adjust for the effects of both of them using only one transformation.

Traditional equipercentile equating is an attempt to adjust the test scores on two forms for their difference in difficulty. But its equated scores are still confounded by the different abilities in the two groups of examinees. Conversely, we could attempt to use a single transformation to adjust the difficulties of any two forms for the differences in abilities between the two groups of examinees that took them. The only thing required seems to be transposing the response matrices for both forms. Their row sums then would represent the total numbers of responses to the items, and division by the number of examinees would give us their \( p \)-values. Continuing the analogy, we would need to apply a Q-Q transformation to the distributions of the \( p \)-values for the two forms. But the “equated \( p \)-values” produced by it still would be confounded by the different difficulties of the items in the two forms.

Obviously, using a single score transformation to address one of the two sources of confounding still leaves us with the confounding caused by the other.

Traditional score equating operates under the assumption that all remaining confounding caused by the ability differences in the two groups is automatically “standardized away” by the introduction of a single target population. But standardizing the effects of confounding is not the same thing as getting rid of them. No matter the choice of target population, the two synthetic score distributions for it still are the results of the different mixtures of the abilities in the two groups of examinees that took X and Y. All earlier issues related to equating bias, group invariance, and lack of equity of observed-score equating are the result of this gratuitous expectation as to the power of standardization.

Conclusion 11: It is impossible to simultaneously adjust test scores for the two main sources of confounding present in any equating study using one transformation.

What Options Are Left?

If the use of a single score transformation does not work, what options are left? Right now, it seems that there are only two: (i) giving up the idea of post hoc score equating altogether and using IRT and (ii) using families of estimated transformations as in local observed-score equating.

IRT

Traditional observed-score equating and IRT share exactly the same goal: accounting for the effects of differences in difficulty and ability between test forms and groups of examinees on test scores. IRT does so by parameterizing both the effects of the ability of each examinee and the difficulty of each item on the level of the individual responses. Consequently, it allows us to ignore the effects of the latter when the interest is in the former, and the other way around (in a specific statistical sense though, which requires the notion of consistent estimators). Observed-score
Some Conceptual Issues in Observed-Score Equating

Equating tries to achieve the same goal but uses a single score transformation on the level of the combination of two fixed test forms and groups of examinees.

Not only is IRT more successful in accounting for the effects of the abilities and difficulties, it is also much more practical. It offers us model-fit statistics to check the assumption of a common variable measured by different sets of items, deals quite naturally with missing data, permits us to change test forms without losing score comparability or even builds them one item at a time as in adaptive testing, and does not assume any random sampling of examinees. All these advantages are supported by a well-developed body of standard statistics.

But its most practical feature might very well be its seamless support of the current transition of educational testing from large-scale, once-a-year administrations of fixed forms for grading or accountability purposes to the use of multiple assessments tailored to more permanent monitoring of students’ progress and instructional feedback to teachers. Fixed-form score equating with its post hoc studies with two fixed groups will have a difficult time supporting this development. But IRT is able to do so in the form of item banking with continuous item writing, field testing, and calibration, in combination with the use of flexible test-assembly algorithms. This setup enables us to follow students’ progress on fixed scales. We are aware of the view of IRT parameter linking as another form of observed-score equating (e.g., Kolen & Brennan, 2004, p. 156). But this linking is something entirely different, namely, the result of a technical requirement—the necessity to work with identifiable models (van der Linden & Barrett, 2013). In the context of item banking, with its option of updating current estimates of field-test item parameters each time new response data arrive, the entire operational item bank serves as an implicit anchor and no post hoc linking is required.

Calibrated item banks even allow us to assemble test forms with number-correct scores automatically equated in the sense of the same indistinguishable observed-score distributions as produced by Equation 6. An example for a retired calibrated pool of 753 items from the Law School Admission Test (LSAT) is given in Figure 11. A reference form of 101 items was assembled from the pool to meet all regular specifications for the test. The remaining pool was used to assemble a new form to meet exactly the same specifications plus two extra requirements: Both its sum of response probabilities,

$$\sum_{i=1}^{101} p_i(\theta),$$  \hspace{1cm} (16)

and sum of their squares,

$$\sum_{i=1}^{101} p_i^2(\theta),$$  \hspace{1cm} (17)

had to be equal to those for the reference test at three typical ability levels ($\theta = -1.2, 0, \text{and } 1.2$). As can be shown in Figure 11, these extra requirements were sufficient to guarantee identical conditional observed-score distributions on the two forms for examinees along the ability scale.
More on this test-assembly solution to observed-score equating and its statistical motivation can be found in van der Linden (2005, Section 5.3) and van der Linden and Luecht (1998). In fact, the same requirements as in Equations 16 and 17 can be imposed on item selection in adaptive testing to automatically equate the number-correct score of each examinee to the same score on a reference form released for score-reporting purposes (van der Linden, 2001). Observe that number-correct scores on an adaptive test equated to those on a common fixed form for each examinee at
Figure 12. Bias functions for the equating of two 40-item test forms using traditional equipercentile and local equating based on IRT ability estimates. The functions are shown for $\theta = -2.0(.5)2.0$.

Local Observed-Score Equating

The second option is statistical estimation of the true transformations in Equation 6. Rather than trying to standardize the confounding effects of the examinees’ abilities out of the equating, this option conditions on them. In doing so, it removes the impact of all other levels than the one of the current examinee from the equating. The result is a different transformation for each ability level, but we already have been made aware of the necessity of this consequence by Conclusion 11.

The true transformations in Equation 6 comprise two different types of quantities: (i) the ability level of the examinee, $\theta$, and (ii) the distributions functions of $X$ and $Y$ given $\theta$. Both of them have to be approximated as closely as possible given the design of the equating study and the test model adopted. The name “local equating” was given to the approach precisely because of this attempt to get as close as possible to the ability level of the examinee and then pool the available information in its neighborhood to approximate his/her conditional distribution functions for the scores on the two forms, the idea being that even a rough approximation already may be better than ignoring the ability levels altogether and pooling data across all test takers in the study.

Depending on the choice of the test model, we may have to estimate either one or both of the two different types of quantities in Equation 6. Two examples illustrating these two possible cases are given. For other examples, we refer to earlier reports in van der Linden (2000, 2001, 2006a, 2006b, 2010, 2011), van der Linden and Wiberg (2010), and Wiberg and van der Linden (2011).

The first example is for an IRT model used to calibrate the items in the two forms. Once the items have been calibrated to sufficient precision, the distributions functions $F_{X|\theta}(x)$ and $G_{Y|\theta}(y)$ do not require any direct estimation; we can calculate them directly from the item parameter estimates using the well-known algorithm introduced in the test-theory literature by Lord and Wingersky (1984). The same thus holds for the true transformation in Equation 6. To prepare an actual equating, these transformations can be calculated for a fine grid of $\theta$ values. The transformation closest to the $\theta$ estimates of the examinees then can be used to find their equated scores on X. Figure 12 shows the bias functions in Equation 10 for the traditional equipercentile
method and this local equating method for an equating from a 40-item test form to another form of the same length. The items in both forms were retired operational items calibrated under the 3PL model from the same pool. The bias functions for the local method were much closer to the ideal of unbiased equating than for the traditional method. In fact, the error shown in the plot for the local method was entirely due to the random error in the examinee’s ability estimates. Consequently, the SEE and MSE functions in Equations 11 and 12 showed essentially the same results, which is why they are omitted here.

The calculations required for the local method in this example are basically the same as for the IRT observed-score equating method in Lord (1980, Chapter 13). The only difference is the mixing of the distributions functions $F_{X|\theta}(x)$ and $G_{Y|\theta}(y)$ across the $\theta$ estimates of all examinees before the calculation of the Q-Q transformation by the latter. The local method omits this mixing operation and calculates the transformation directly for each $\theta$ estimate. The difference in success due to the omission of this single step is remarkable, though. Also, observe that the current local method can be used with any equating design that allows us to produce estimates of the $\theta$ parameters for the test takers. An example with $\theta$s estimated from an anchor test was presented by Janssen, Magis, San Martin, and Del Pino (2009).

The second example adopts the classical test-theory model for an anchor test design. For this choice, we need statistical estimates both of the abilities of the examinees and the conditional observed-score distributions given their abilities. The presence of an anchor test helps us with both. Let $A$ denote the anchor test and $A$ the observed score on it for an arbitrary examinee. The approach relies on the assumption of $X$, $Y$, and $A$ measuring a common variable. Let $\theta$ denote this variable, measured on a scale that need not be known. It holds that the classical true score on $A$, denoted as $\tau(A)$, is a monotone function of $\theta$. Because of this monotone relationship, conditioning the observed scores on $X$ and $Y$ on $\theta$ gives the same result as conditioning them on $\tau(A) = \tau$. However, we have an estimate of $\tau$ in the form of the observed score $A = a$. Thus, all we have to do is to estimate

$$
\varphi(y, a) = F_{X|a}^{-1}(G_{Y|a}(x))
$$

(18)
directly from the data available in the equating study. In earlier reports, we have used the term “proxy” for this role of $A$ in the estimation of the true equating transformations. Obviously, $A = a$ is a hopeless estimate of the unknown (even incompletely defined) $\theta$. But it is an excellent estimate of a monotone function of it and therefore serves as an obvious substitute to condition on. Exploration of other possibly useful types of proxies (multiple regressors in the form of reliable background variables; response times; earlier test scores?) seems an attractive option for observed-score equating research to pursue.

Figure 13 shows the results from a study with the same two 40-item forms as in the previous example along with an anchor test of the same length. The number-correct scores for the examinees were calculated from the responses generated under the 3PL model for these three forms. (Just to make sure, the model itself was not used in the equating; it was used only to generate the responses.) The results for the local method in Equation 18 were compared with those for traditional chain and poststratification equating for an anchor test design. Again, the bias functions for
the local method were much closer to the ideal of unbiased equating than for the two traditional methods. The remaining error in the plot for the local method was basically random error due to the use of the observed anchor-test scores as estimates of their true scores.

Again, the computational efforts required for the local method are less than for the two others. Unlike chain equating, there is no need to go from Y to A using the distributions for one population and then from A to X for another. Neither do we have to reweigh the conditional distributions of the score on X and Y given \( A = a \), as in poststratification equating. The only step this local method requires is using the observed distributions of X and Y given \( A = a \) to calculate an estimate of Equation 18.

This is a good moment to be more precise about the roles of the test length and the number of examinees as determinants of equating error discussed earlier (Conclusion 7). For both local methods used in our two examples, the role of test length is critical. For the first method, equating error depends on the length of Y from which \( \hat{\theta} \) is estimated (or the length of the anchor test in the version of the method used in Janssen et al., 2009). If we increase the length of this form, \( \hat{\theta} \) converges to \( \theta \) for each examinee and all equating error vanishes. Actually, it is the statistical information about \( \theta \) in the test that counts; equating error also vanishes with an increase in the discriminating power of its items. Likewise, for the local method in the second example, an increase in the length of the anchor test or the discriminating power of its items makes \( A = a \) a more accurate estimate of its true score and hence a better proxy of \( \theta \). Under these conditions, equating error vanishes as well. For empirical examples of these effects, we refer to van der Linden (2006a) and van der Linden and Wiberg (2010).

The second method, however, faces us with a trade-off between the length of the anchor test and numbers of examinees: For a fixed total number of examinees, any
increase in the length of the anchor test, although a desirable feature by itself, leads to a decrease in the number of examinees per possible score on it and therefore less accurate estimation of the distributions of the score on X and Y given \( A = a \). At this point, the power of the smoothing and continuizing techniques developed for kernel equating (von Davier, Holland, & Thayer, 2004) may come in handy. Studies looking into the possibilities of combining a local approach with these techniques are already under way (Wiberg, van der Linden, & von Davier, in press).

Concluding Remarks

One of my motives for writing this discussion article has been a concern about the legacy of the earlier work on score equating by Frederic Lord, culminating in the analysis in his 1980 monograph. I have always wondered why his seminal early work has not had any serious follow-up. In fact, Lord himself made a remarkable change shortly after publication of the monograph. In his article on the standard error of equipercentile equating (Lord, 1982), he shifted his interest from the groups of examinees that actually took the two test forms to “some specified population,” exchanged measurement error for sampling error, no longer used any test theory, downplayed the possibility of equating bias, and in doing so may actually have contributed to the relative oblivion of his earlier work.

Nevertheless, in order to get observed-score equating on better footing, a revaluation of Lord’s criterion of equity is required. The only necessary additional step is the insight that rather than a single transformation for all examinees (an assumption automatically adopted since the early days of observed-score equating), the criterion involves a different transformation for each of their ability levels. At first sight, the step may seem to face us with several new challenges. But the only thing required is conditioning of the observed-score distributions on well-chosen ability estimates or proxies for them and evaluating the results using standard statistical criteria as in Equations 10–12. The empirical examples in this article illustrate just a few of the possibilities.

Educational measurement has gone through exactly the same process of replacing a one-size-fits-all estimate with ability-specific estimates before, for exactly the same reason. For a long time, it was common to use a single standard error of measurement for every examinee that took the same test until we became aware of how wrong this actually was: Measurement error strongly depends on the ability of the examinee and it tends to be larger at high or low levels of ability than more toward the middle of the scale. If we combine estimates of it for different levels, the result becomes both group-dependent and biased. Nearly every testing program now reports estimated standard errors of measurement conditional on ability. It is time for observed-score equating to go through the same transition.

It may be interesting to know that the distinction between equivalence for a group and each of its individual exists in other fields as well. One of the main problems in biostatistics is how to prove the equivalence of different treatments. A well-known example of this bioequivalence problem is the difference between treatments with brand and generic drugs. Of course, this problem does not involve any equating of scores, but the discussion of the criteria of what constitutes bioequivalence is
relevant. The three common definitions are: population bioequivalence (same distributions of the effects of the treatments in a population of patients), average bioequivalence (equal mean effects in a population), and individual bioequivalence (same effects for each individual patient) (e.g., Chow, 1997; Berger & Hsu, 1996). The first definition is analogous to what the tradition of observed-score equating has been after so far. The second corresponds to the idea of mean equating, a stronger version of equating than linear equating, which assumes score distributions differing only in their means. The last definition is the most strict. As the effects of medical treatments have a random component, it requires effect distributions of the treatments for each individual patient that are indistinguishable from each other—the pendant of Lord’s criterion of equity. Is that not what we expect when we buy a generic drug?

Acknowledgments

The author is indebted to Soe Aung and Joanna Tomkowicz for their assistance with some of the empirical examples in this article, and to the Law School Admission Council (LSAC) for a data set used in this study. The opinions and conclusions contained in this article are those of the author and do not necessarily reflect the policy and position of the LSAC.

References


Luce, R. D. (1986). *Response times: Their roles in inferring elementary mental organization*. Oxford, UK: Oxford University Press.


Author

WIM J. VAN DER LINDEN is Chief Research Scientist, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940; wim-vanderlinden@ctb.com. His primary research interests include test theory, applied statistics, and research methods.